EMPLOYEE CHURN PREDICTION

USING PYTHON

A Project Report

submitted in partial fulfillment of the requirements

of

AI and ML fundamentals with Cloud Computing and Gen AI

by

RAHAMATH NISHA M (aut8139001)

Under the Guidance of

P. Raja

Master Trainer

ACKNOWLEDGEMENT

First, we wish to express our gratitude to the almighty God, the lifter of our head and the giver of all wisdom. He truly deserves all the glory and the honour, for he did not let our foot slip.

I would like to express our deep sense of gratefulness to **Mr. P.RAJA** Master trainer, Tech Saksham for his guidance in all activity and playing a major role for successful completion of this project.

I really grateful to **Mr. T.CHRISTHURAJ, M.E.,** Naan Mudhalvan Coordinator for advising me and helped me to complete my project easily and affectively on time.

ABSTRACT

Employees are the valuable assets of any company. But if they quit jobs unexpectedly, it may incur huge cost to any company. Because new hiring will consume not only money and time but also the freshly hired employees take time to make the respective company profitable. In the past, most of the focus on the 'rates' such as attrition rate and retention rates. HR Managers compute the previous rates try to predict the future rates using data warehousing tools. Hence, in this study, I try to build a model employee churn prediction model which predicts either the employees will leave their current company or stay in the company based on Employee churn dataset obtained from Kaggle website. Within the study's scope, we have trained standard and sequential models using python then evaluated the model performance. In the experimental part, the histogram is generated, which shows the contrast between left employees vs. salary, department, satisfaction level, etc. Here, you can predict who, and when an employee will terminate the service. Employee churn is expensive, and incremental improvements will give significant results. It will help us in designing better retention plans and improving employee satisfaction.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 PROBLEM STATEMENT:

Analyze employee churn. Find out why employees are leaving the company, and learn to predict who will leave the company. Employee churn can be defined as a leak or departure of an intellectual asset from a company or organization. Alternatively, in simple words, you can say, when employees leave the organization is known as churn. Another definition can be when a member of a population leaves a population, is known as churn.

1.2 MOTIVATION:

In the past, most of the focus on the 'rates' such as attrition rate and retention rates. HR Managers compute the previous rates try to predict the future rates using data warehousing tools. These rates present the aggregate impact of churn, but this is the half picture. Another approach can be the focus on individual records in addition to aggregate. There are lots of case studies on customer churn are available. In customer churn, you can predict who and when a customer will stop buying. Employee churn is similar to customer churn. It mainly focuses on the employee rather than the customer. Here, you can predict who, and when an employee will terminate the service. Employee churn is expensive, and incremental improvements will give significant results. It will help us in designing better retention plans and improving employee satisfaction.

In Research, it was found that employee churn will be affected by age, tenure, pay, job satisfaction, salary, working conditions, growth potential and employee's perceptions of fairness. Some other variables such as age, gender, ethnicity, education, and marital status, were essential factors in the prediction of employee churn. In some cases, such as the employee with niche skills are harder to replace. It affects the ongoing work and productivity of existing employees.

Acquiring new employees as a replacement has its costs such as hiring costs and training costs. Also, the new employee will take time to learn skills at the similar level of technical or business expertise knowledge of an older employee. Organizations tackle this problem by applying machine learning techniques to predict employee churn, which helps them in taking necessary actions.

Following points help you to understand, employee and customer churn in a better way:

- Business chooses the employee to hire someone while in marketing you don't get to choose your customers.
- Employees will be the face of your company, and collectively, the employees produce everything your company does.
- Losing a customer affects revenues and brand image. Acquiring new customers is difficult and costly compared to retain the existing customer. Employee churn also painful for companies an organization. It requires time and effort in finding and training a replacement.

Employee churn has unique dynamics compared to customer churn. It helps us in designing better employee retention plans and improving employee satisfaction. Data science algorithms can predict the future churn.

1.3 OBJECTIVE:

- To analyze the employee churn using Dataset downloaded from Kaggle website.
- To analyze the employee churn based on age, tenure, pay, job satisfaction, salary, working conditions, growth potential and employee's perceptions of fairness. Some other variables such as age, gender, ethnicity, education, and marital status, were essential factors in the prediction of employee churn.
- To train the model using machine learning algorithms to find out whether an employee leave the company or not.
- Evaluation of the model performance.
- Designing better employee retention plans and improving employee satisfaction using predictive model.

1.4 SCOPE OF THE PROJECT:

Employee churn is similar to customer churn. It mainly focuses on the employee rather than the customer. Here, you can predict who, and when an employee will terminate the service. Employee churn is expensive, and incremental improvements will give significant results. It will help us in designing better retention plans and improving employee satisfaction. In Research, it was found that employee churn will be affected by age, tenure, pay, job satisfaction, salary, working conditions, growth potential and employee's perceptions of fairness. Some other variables such as age, gender, ethnicity, education, and marital status, were essential factors in the prediction of employee churn. In some cases, such as the employee with niche skills are harder to replace. It affects the ongoing work and productivity of existing employees. Acquiring new employees as a replacement has its costs such as hiring costs and training costs. Also, the new employee will take time to learn skills at the similar level of technical or business expertise knowledge of an older employee. Organizations tackle this problem by applying machine learning techniques to predict employee churn, which helps them in taking necessary actions. Employee churn has unique dynamics compared to customer churn. It helps us in designing better employee retention plans and improving employee satisfaction. Data science algorithms can predict the future churn.

CHAPTER 2

LITERATURE SURVEY

2.1 Review relevant literature or previous work in this domain.

Nowadays, in order to assess the employee performances in organizations, organizations are considering reward schemes. This related work highlights the idea that if the organization offered a smaller number of rewards schemes, it would be difficult to retain employees. A study conducted an online survey for collecting the data from companies; they applied regression and correlation statistical methods on survey data to obtain the essential facts for the IT sectors. They identify the essential features that contribute to employee turnover using the SPSS tool. In order to analyze churn factors, they conducted an online survey which is based on a questionnaire. Organizational commitment is a hidden relationship between organizations and employees. To sustain the skilled employees in a company, the most crucial role is that of managers. Their study has three organizational commitments, and they applied mathematical methods for determining the hidden relation between organizational commitment and intention of employee turnover. This related study helps HR managers take proactive steps and create good employee policies. Another related study determines the essential factors as satisfaction level, workload, and career opportunity. It also suggests that by providing a friendly atmosphere, low burden of work, and increased career opportunities, there will be increased employee retention and a decrease in the churn rate. They also recommend that salaries are not a practical tool to hold employees in organizations. Some of the analytical approaches applied in this area are presented below. The research mentioned in aims to avoid this negative impact; therefore, their study builds the prediction model to predict future churners. For the comparison and evaluation of algorithms, the research implements the three algorithms and compares them. Employee churn is a severe problem in organizations. If the organization needs economic benefit, then there is a need to minimize attrition.

2.3 Existing models, techniques, or methodologies related to the problem.

1. Weeramanthrie, T.T.; Thilakumara, C.N.; Wijesiri, K.N.A.C.; Fernando, N.I.; Thelijjagoda, S.; Gamage, A. ARROW: A web-based employee turnover analysis tool for effective human resource management in large-scale organizations. In Proceedings of the National Information Technology Conference (NITC), Columbo, Sri Lanka, 14–15 September 2017; pp. 136–140.
2. Dolatabadi, S.H.; Keynia, F. Designing of customer and employee churn prediction model based on data mining method and neural predictor. In Proceedings of the 2nd International Conference on Computer and Communication Systems (ICCCS), Kraków, Poland, 11–14 July 2017; pp. 74–77.
3. Sethunga, S.; Perera, I. Impact of Performance Rewards on Employee Turnover in Sri Lankan IT Industry. In Proceedings of the Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, 30 May–1 June 2018; pp. 114–119.
4. Sisodia, D.S.; Vishwakarma, S.; Pujahari, A. Evaluation of machine learning models for employee churn prediction. In Proceedings of the International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, 23 –24 November 2017; pp. 1016–1020.
5. Srivastava, D.K.; Nair, P. Employee attrition analysis using predictive techniques. In Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems, Ahmedabad, India, 25–26 March 2017; pp. 293–300.
6. Saghir, M.; Bibi, Z.; Bashir, S.; Khan, F.H. Churn prediction using neural network based individual and ensemble models. In Proceedings of the 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Bhurban, Pakistan, 8–12 January 2019; pp. 634–639.
7. Esmaieeli Sikaroudi, A.M.; Ghousi, R.; Sikaroudi, A. A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). *J. Ind. Syst. Eng.* 2015, *8*, 106–121.

CHAPTER 3

PROPOSED METHODOLOGY

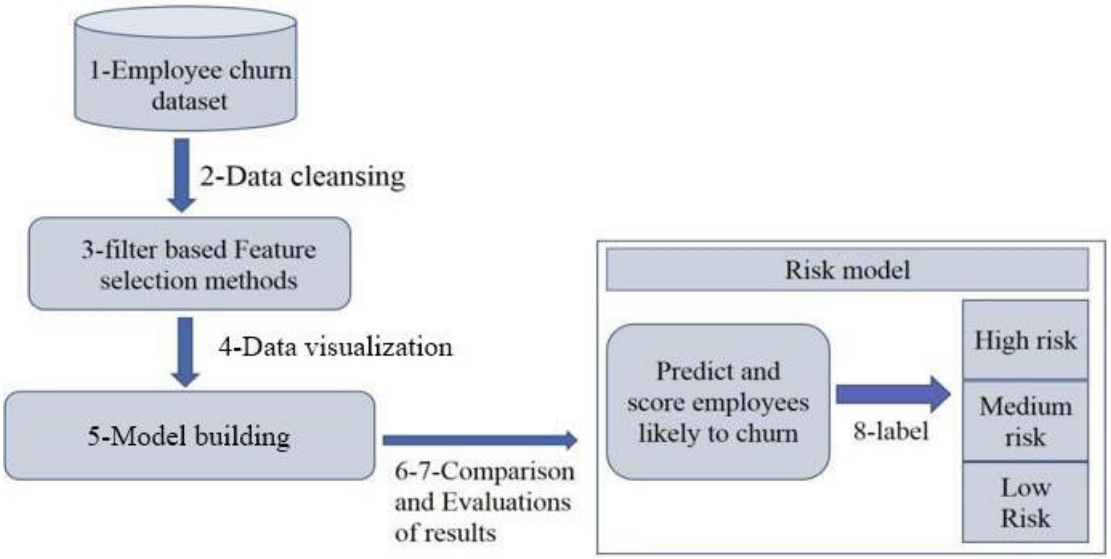


Figure 1. Employee churn analysis flow diagram of software industry.

1. We have used an employee churn analytics HR dataset that consists of both types of employees, churn and non-churn.
2. Data selection and preprocessing are performed in the second step.
3. In the third step, we have used four filter-based methods.
4. We have selected the top N highest-ranking variables from these four methods.
5. After selecting the top features, this research applies five ML algorithms.
6. Next, we compared the performance parameters (accuracy, precision and recall) of five ML algorithms and evaluated the classification results after splitting the dataset into different ratios with N-ranked features.
7. We chose the best algorithm with a filter-based method for building the risk model.
8. Finally, we created a risk model for non-churn employees in order to further classify them into the different risk zones.

3.1 EXPLORATORY ANALYSIS

Exploratory Data Analysis is an initial process of analysis, in which you can summarize characteristics of data such as pattern, trends, outliers, and hypothesis testing using descriptive statistics and visualization.

IMPORTING MODULES

```
[1]: import numpy as np
     import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
     %matplotlib inline
```

LOADING DATASET

Let's first load the required HR dataset using pandas read CSV function.

```
[2]: data=pd.read_csv('employee churn.csv')
```

We can take a closer look at the data took help of "head()"function of pandas library which returns first five observations.

```
[3]: data.head()
```

[3]:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	Departments	salary
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.80	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.72	0.87	5	223	5	0	1	0	sales	low
4	0.37	0.52	2	159	3	0	1	0	sales	low

We can take a closer look at the data took help of "tail()"function of pandas library which returns first five observations.

```
[4]: data.tail()
```

[4]:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	Departments	salary
14994	0.40	0.57	2	151	3	0	1	0	support	low
14995	0.37	0.48	2	160	3	0	1	0	support	low
14996	0.37	0.53	2	143	3	0	1	0	support	low
14997	0.11	0.96	6	280	4	0	1	0	support	low
14998	0.37	0.52	2	158	3	0	1	0	support	low

We can check attributes names and datatypes using info().

```
[5]: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   satisfaction_level     14999 non-null  float64
 1   last_evaluation        14999 non-null  float64
 2   number_project         14999 non-null  int64
 3   average_montly_hours   14999 non-null  int64
 4   time_spend_company     14999 non-null  int64
 5   Work_accident          14999 non-null  int64
 6   left                   14999 non-null  int64
 7   promotion_last_5years  14999 non-null  int64
 8   Departments            14999 non-null  object
 9   salary                 14999 non-null  object
dtypes: float64(2), int64(6), object(2)
memory usage: 1.1+ MB
```

- This dataset has 14,999 samples, and 10 attributes (6 integer, 2 float, and 2 objects).
- No variable column has null/missing values.

DATA INSIGHTS

In the given dataset, you have two types of employee one who stayed and another who left the company. So, you can divide data into two groups and compare their characteristics. Here, you can find the count of both the groups using groupby() and count() function.

```
data.groupby(['left']).count()
```

left	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years	Departments	salary
0	11428	11428	11428	11428	11428	11428	11428	11428	11428
1	3571	3571	3571	3571	3571	3571	3571	3571	3571

The describe() function in pandas is convenient in getting various summary statistics. This function returns the count, mean, standard deviation, minimum and maximum values and the quantiles of the data.

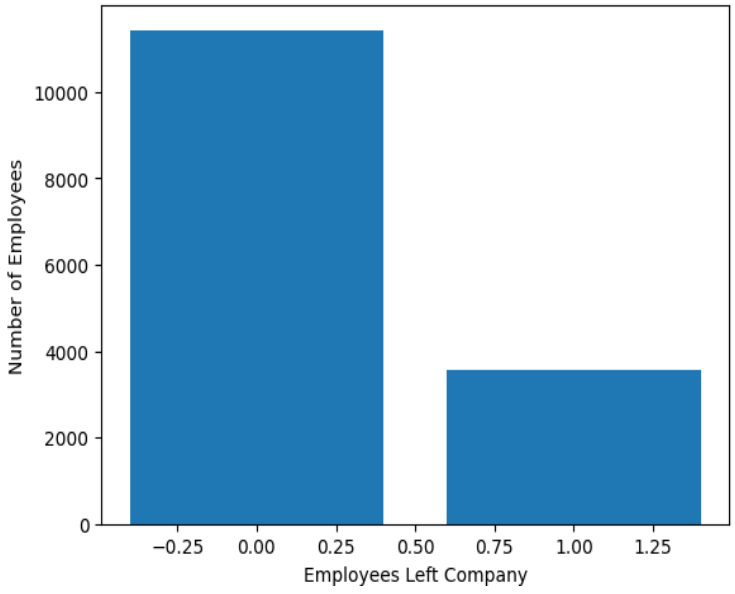
```
data.describe()
```

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years
count	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000
mean	0.612834	0.716102	3.803054	201.050337	3.498233	0.144610	0.238083	0.021268
std	0.248631	0.171169	1.232592	49.943099	1.460136	0.351719	0.425924	0.144281
min	0.090000	0.360000	2.000000	96.000000	2.000000	0.000000	0.000000	0.000000
25%	0.440000	0.560000	3.000000	156.000000	3.000000	0.000000	0.000000	0.000000
50%	0.640000	0.720000	4.000000	200.000000	3.000000	0.000000	0.000000	0.000000
75%	0.820000	0.870000	5.000000	245.000000	4.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	7.000000	310.000000	10.000000	1.000000	1.000000	1.000000

3.2 DATA VISUALIZATION

EMPLOYEES LEFT

```
left_count=data.groupby('left').count()
plt.bar(left_count.index.values, left_count['satisfaction_level'])
plt.xlabel('Employees Left Company')
plt.ylabel('Number of Employees')
plt.show()
```



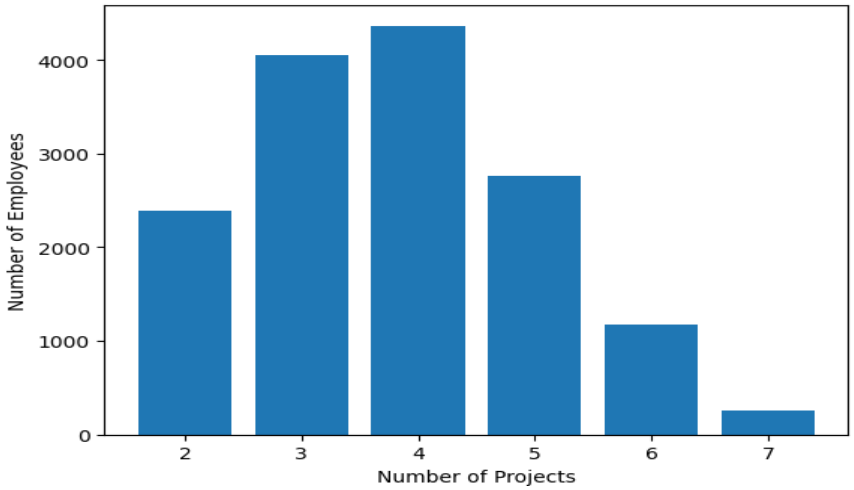

```
data.left.value_counts()
```

```
left
0    11428
1     3571
Name: count, dtype: int64
```

Here, we can see out of 15,000 approx. 3,571 were left, and 11,428 stayed. The no of employee left is 23 % of the total employment.

NUMBER OF PROJECTS

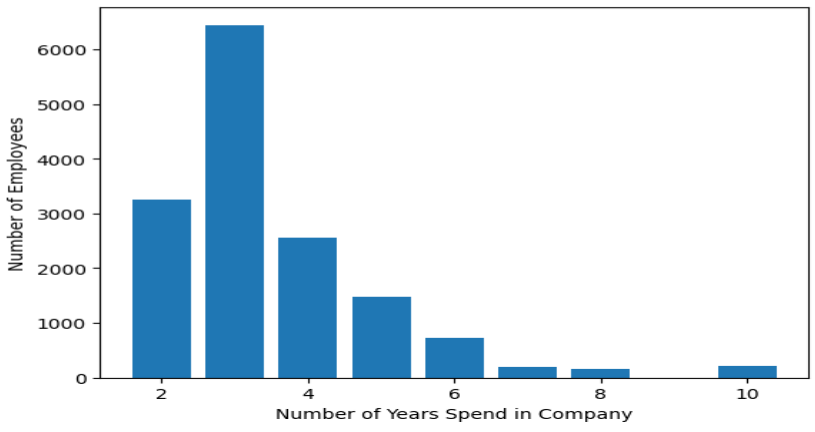
```
num_projects=data.groupby('number_project').count()
plt.bar(num_projects.index.values, num_projects['satisfaction_level'])
plt.xlabel('Number of Projects')
plt.ylabel('Number of Employees')
plt.show()
```



Most of the employee is doing the project from 3-5.

TIME SPENT IN COMPANY

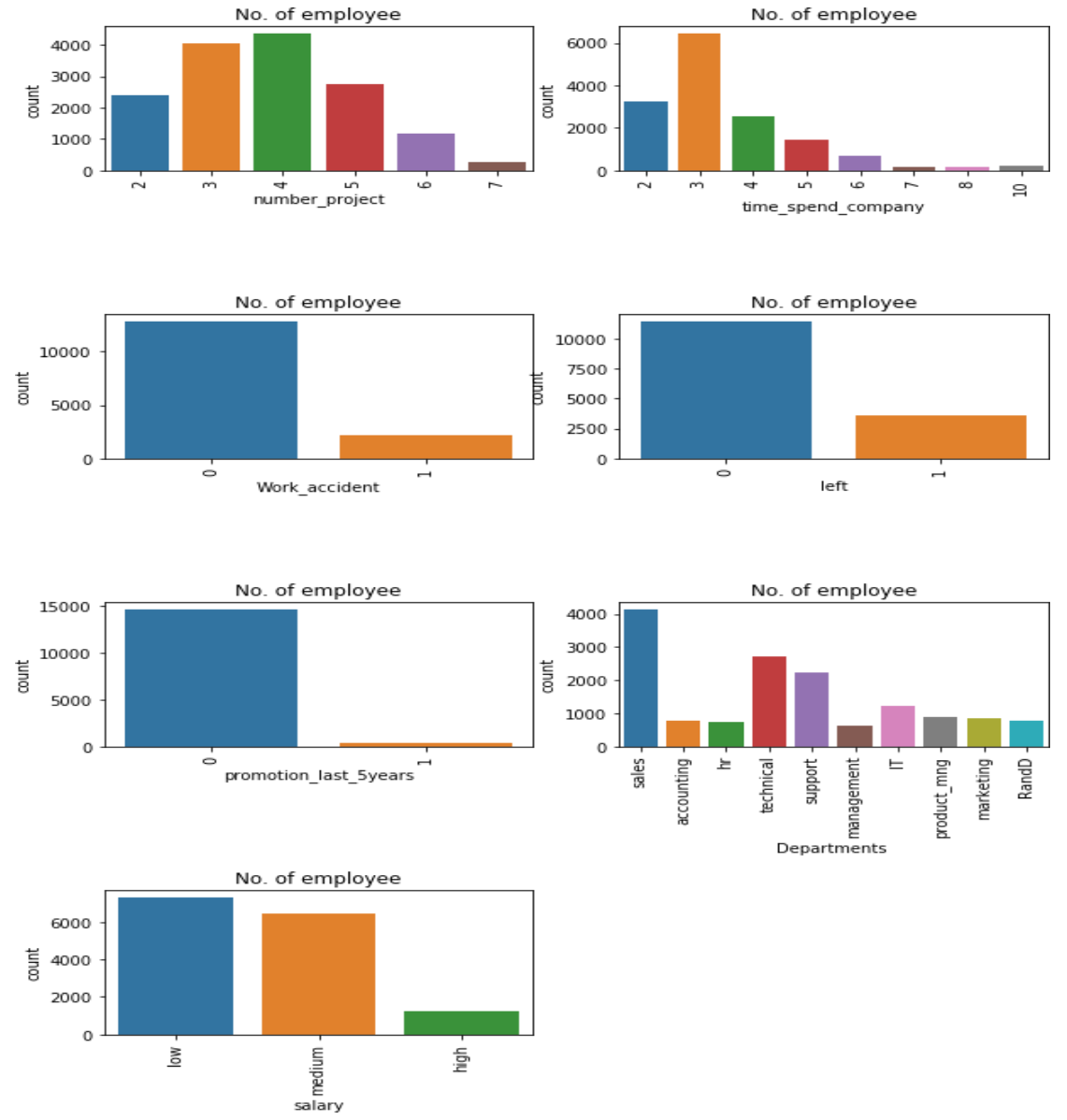
```
time_spent=data.groupby('time_spend_company').count()
plt.bar(time_spent.index.values, time_spent['satisfaction_level'])
plt.xlabel('Number of Years Spend in Company')
plt.ylabel('Number of Employees')
plt.show()
```



Most of the employee experience between 2-4 years. Also, there is a massive gap between 3 years and 4 years experienced employee.

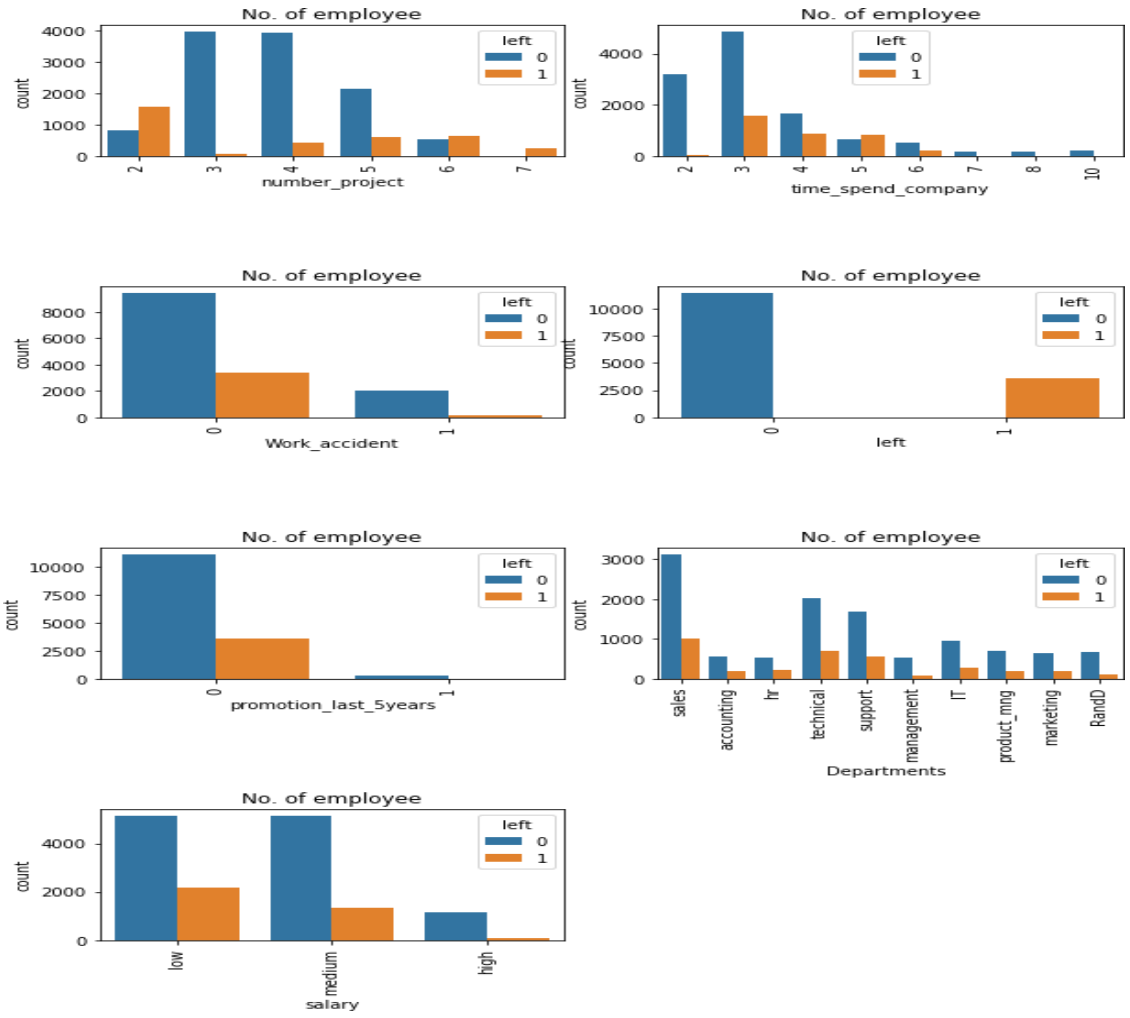
SUBPLOTS USING SEABORN

```
features=['number_project','time_spend_company','Work_accident','left', 'promotion_last_5years','Departments ','salary']
fig=plt.subplots(figsize=(10,15))
for i, j in enumerate(features):
    plt.subplot(4, 2, i+1)
    plt.subplots_adjust(hspace = 1.0)
    sns.countplot(x=j,data = data)
    plt.xticks(rotation=90)
    plt.title("No. of employee")
```



- Most of the employee is doing the project from 3-5.
- There is a huge drop between 3 years and 4 years experienced employee.
- The no of employee left is 23 % of the total employment.
- A decidedly a smaller number of employees get the promotion in the last 5 year.
- The sales department is having maximum number of employees followed by technical and support.
- Most of the employees are getting salary either medium or low.

```
fig=plt.subplots(figsize=(10,15))
for i, j in enumerate(features):
    plt.subplot(4, 2, i+1)
    plt.subplots_adjust(hspace = 1.0)
    sns.countplot(x=j,data = data, hue='left')
    plt.xticks(rotation=90)
    plt.title("No. of employee")
```



- Those employees who have the number of projects more than 5 were left the company.
- The employee who had done 6 and 7 projects, left the company it seems to like that they were overloaded with work.
- The employee with five-year experience is leaving more because of no promotions in last 5 years and more than 6 year experience are not leaving because of affection with the company.
- Those who promotion in last 5 years they didn't leave, i.e., all those left they didn't get the promotion in the previous 5 years.

DATA VISUALIZATION SUMMARY

Following features are most influencing a person to leave the company:

- **Promotions:**

  Employees are far more likely to quit their job if they haven't received a promotion in the last 5 years.

- **Time with Company:**

  Here, the three-year mark looks like a time to be a crucial point in an employee's career. Most of them quit their job around the three-year mark. Another important point is 6-years point, where the employee is very unlikely to leave.

- **Number of Projects:**

  Employee engagement is another critical factor to influence the employee to leave the company. Employees with 3-5 projects are less likely to leave the company. The employee with less and a greater number of projects are likely to leave.

- **Salary:**

  Most of the employees that quit among the mid or low salary groups.

3.3 CLUSTER ANALYSIS

```
from sklearn.cluster import KMeans
left_emp =  data[['satisfaction_level', 'last_evaluation']][data.left == 1]
kmeans = KMeans(n_clusters = 3, random_state = 0).fit(left_emp)
```

```
left_emp['label'] = kmeans.labels_
plt.scatter(left_emp['satisfaction_level'], left_emp['last_evaluation'], c=left_emp['label'],cmap='Accent')
plt.xlabel('Satisfaction Level')
plt.ylabel('Last Evaluation')
plt.title('3 Clusters of employees who left')
plt.show()
```

- High Satisfaction and High Evaluation (Shaded by green color in the graph), you can also call them Winners.
- Low Satisfaction and High Evaluation (Shaded by blue color in the graph), you can also call them Frustrated.
- Moderate Satisfaction and moderate Evaluation (Shaded by grey color in the graph), you can also call them 'Bad match'

CHAPTER 4

IMPLEMENTATION AND RESULT

4.1 BUILDING A PREDICTION MODEL

PRE-PROCESSING DATA

Lots of machine learning algorithms require numerical input data, so we need to represent categorical columns in a numerical column.

In order to encode this data, we could map each value to a number. e.g. Salary column's value can be represented as low:0, medium:1, and high:2.

This process is known as label encoding, and sklearn conveniently will do this for using LabelEncoder.

```
# Import LabelEncoder
from sklearn import preprocessing
#creating labelEncoder
le = preprocessing.LabelEncoder()
# Converting string labels into numbers.
data['salary']=le.fit_transform(data['salary'])
data['Departments ']=le.fit_transform(data['Departments '])
```

Here, we imported preprocessing module and created Label Encoder object. Using this LabelEncoder object you fit and transform "salary" and "Departments " column into numeric column.

SPLIT TRAIN AND TEST DATA

To understand model performance, dividing the dataset into a training set and a test set is a good strategy.

Let's split dataset by using function train_test_split(). You need to pass 3 parameters features, target, and test_set size. Additionally, you can use random_state to select records randomly.

```
#Spliting data into Feature
X=data[['satisfaction_level', 'last_evaluation', 'number_project',
       'average_montly_hours', 'time_spend_company', 'Work_accident',
       'promotion_last_5years', 'Departments ', 'salary']]
y=data['left']
```

```
# Import train_test_split function
from sklearn.model_selection import train_test_split

# Split dataset into training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=42)  # 80% training and 20% test
```

Here, Dataset is broken into two parts in ratio of 80:20. It means 80% data will used for model training and 20% for model testing.

MODEL BUILDING

Let's build an employee churn prediction model.

Here, we are going to predict churn using Gradient Boosting Classifier.

First, import the GradientBoostingClassifier module and create Gradient Boosting classifier object using GradientBoostingClassifier() function.

Then, fit your model on train set using fit() and perform prediction on the test set using predict().

```
#Import Gradient Boosting Classifier model
from sklearn.ensemble import GradientBoostingClassifier

#Create Gradient Boosting Classifier
gb = GradientBoostingClassifier()

#Train the model using the training sets
gb.fit(X_train, y_train)

#Predict the response for test dataset
y_pred = gb.predict(X_test)
```

4.2 EVALUATING MODEL PERFORMANCE

```
#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
# Model Precision
print("Precision:",metrics.precision_score(y_test, y_pred))
# Model Recall
print("Recall:",metrics.recall_score(y_test, y_pred))
```

```
Accuracy: 0.9736666666666667
Precision: 0.9617083946980854
Recall: 0.9249291784702549
```

Well, we got a classification rate of 97%, considered as good accuracy.

Precision: Precision is about being precise, i.e., how precise your model is. In other words, you can say, when a model makes a prediction, how often it is correct. In your prediction case, when your Gradient Boosting model predicted an employee is going to leave, that employee actually left 96% of the time.

Recall: If there is an employee who left present in the test set and your Gradient Boosting model can identify it 92% of the time.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 KEY FINDINGS

- From the data set we can see out of 15,000 approx. 3,571 were left, and 11,428 stayed. The no of employee left is 23 % of the total employment.
- Most of the employee is doing the project from 3-5.
- Those employees who have the number of projects more than 5 were left the company.
- The employee who had done 6 and 7 projects, left the company it seems to like that they were overloaded with work.
- The employee with five-year experience is leaving more because of no promotions in last 5 years and more than 6 year experience are not leaving because of affection with the company.
- Those who promotion in last 5 years they didn't leave, i.e., all those left they didn't get the promotion in the previous 5 years.

Following features are most influencing a person to leave the company:

- **PROMOTIONS:**

  Employees are far more likely to quit their job if they haven't received a promotion in the last 5 years.

- **TIME WITH COMPANY:**

  Here, the three-year mark looks like a time to be a crucial point in an employee's career. Most of them quit their job around the three-year mark. Another important point is 6-years point, where the employee is very unlikely to leave.

- **NUMBER OF PROJECTS:**

  Employee engagement is another critical factor to influence the employee to leave the company. Employees with 3-5 projects are less likely to leave the company. The employee with less and a greater number of projects are likely to leave.

- **SALARY:**

  Most of the employees that quit among the mid or low salary groups.

5.2 GITHUB LINK OF THE PROJECT

https://github.com/Rahamathnisha1011/EMPLOYEE-CHURN-PREDICTION-USING-PYTHON.git

5.3 VIDEO RECORDING OF THE PROJECT DEMONSTRATION

https://drive.google.com/file/d/1YoAucz_VZxKKvqIT0mCYq6aqvuBCNNpd/view?usp=drivesdk

5.4 CONCLUSION

Organizations lose money, time, and effort as a result of employee churn. A trained and experienced person is difficult and expensive to replace; thus this is a major problem. In order to forecast future employee turnover and understand its causes, we examine data on both past and present employees. The findings of this study show that data mining techniques can be applied to create trustworthy and precise forecast models for employee churn. Distinguishing churners from non-churners is only one aspect of the churn prediction challenge. By applying exploratory data analysis and data mining techniques, we can predict the probability of each employee leaving their job and assign them a score to enable them to develop retention strategies. In this research, we determine that using GradientBoostingClassifier() function. The model performance shows 97% accuracy and 96% precision. The prediction model shows that the most influential factors are satisfaction level, the number of projects, time spent on the company, and last evaluation.

After the analysis of the classification results, this employee churn prediction model supports the organizations in how to retain valuable employees and to make better decisions. The work's limitations in this research include only using filter-based methods, which are computationally efficient, but if we want more accurate results of our ML models we could use embedded methods, which are costly for computation. For a future direction we plan to build hybrid-based methods which are a combination of filter- and embedded-based methods in order to build more accurate and comprehensive models for organizations to utilize for the betterment of the employees and future prospects

REFERENCES

1. Dolatabadi, S.H.; Keynia, F. Designing of customer and employee churn prediction model based on data mining method and neural predictor. In Proceedings of the 2nd International Conference on Computer and Communication Systems (ICCCS), Kraków, Poland, 11–14 July 2017; pp. 74–77.
2. Weeramanthrie, T.T.; Thilakumara, C.N.; Wijesiri, K.N.A.C.; Fernando, N.I.; Thelijjagoda, S.; Gamage, A. ARROW: A web-based employee turnover analysis tool for effective human resource management in large-scale organizations. In Proceedings of the National Information Technology Conference (NITC), Columbo, Sri Lanka, 14–15 September 2017; pp. 136–140.
3. Sethunga, S.; Perera, I. Impact of Performance Rewards on Employee Turnover in Sri Lankan IT Industry. In Proceedings of the Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, 30 May–1 June 2018; pp. 114–119.
4. Wei, G.U.O.; Tai, L.I. An empirical study on organizational commitment and turnover of its industry. In Proceedings of the International Conference on E-Business and E-Government, Guangzhou, China, 7–9 May 2010; pp. 904–906.
5. David, S.; Kaushik, S.; Verma, H.; Sharma, S. Attrition in "IT" Sector. *Int. J. Core Eng. Manag. IJCEM* 2015, *2*, 74–92.
6. Mura, L.; Zsigmond, T.; Machová, R. The effects of emotional intelligence and ethics of SME employees on knowledge sharing in Central-European countries. *Oeconomia Copernic.* 2021, *12*, 907–934.
7. Szeiner, Z.; Kovács, Á.; Zsigmond, T.; Mura, L.; Sanders, E.; Poor, J. An empirical study of consulting in a transitional economy in the Central European region during COVID-19. *J. East. Eur. Cent. Asian Res. (JEECAR)* 2022, *9*, 471–485.
8. Alamsyah, A.; Salma, N. A Comparative Study of Employee Churn Prediction Model. In Proceedings of the 4th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 7–8 August 2018; pp. 1–4.
9. Carraher, S.M. Turnover prediction using attitudes towards benefits, pay, and pay satisfaction among employees and entrepreneurs in Estonia, Latvia, and Lithuania. *Balt. J. Manag.* 2011, *6*, 25–52.

10. Yiğit, İ.O.; Shourabizadeh, H. An approach for predicting employee churn by using data mining. In Proceedings of the International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 16–17 September 2017; pp.
11. Cheripelli, R.; Ajitha, P.V. Evaluation of Machine Learning Models for Employee Churn Prediction. *Solid State Technol.* 2020, *63*, 2482–2487.
12. Sisodia, D.S.; Vishwakarma, S.; Pujahari, A. Evaluation of machine learning models for employee churn prediction. In Proceedings of the International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, 23–24 November 2017; pp. 1016–1020.
13. Jain, H.; Yadav, G.; Manoov, R. Churn Prediction and Retention in Banking, Telecom and IT Sectors Using Machine Learning Techniques. In *Advances in Machine Learning and Computational Intelligence*, 1st ed.; Springer: Singapore, 2021; Volume 1, pp. 137–156.
14. Yeom, S.; Giacomelli, I.; Fredrikson, M.; Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In Proceedings of the 31st Computer Security Foundations Symposium (CSF), Oxford, UK, 9–12 July 2018; pp. 268–282.
15. Ghosh, P.; Satyawadi, R.; Joshi, J.P.; Shadman, M. Who stays with you? Factors predicting employees' intention to stay. *Int. J. Organ. Anal.* 2013, *21*, 288–312.

SOURCE CODE

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
data=pd.read_csv('employee churn.csv')
data.head()
data.tail()
data.info()
data.describe()
data.shape
data.dtypes
pd.isna(data).sum()
data.isna().count()
data.groupby(['left']).count()
left_count=data.groupby('left').count()
plt.bar(left_count.index.values, left_count['satisfaction_level'])
plt.xlabel('Employees Left Company') plt.ylabel('Number of Employees')
plt.show()
data.left.value_counts()
num_projects=data.groupby('number_project').count()
plt.bar(num_projects.index.values, num_projects['satisfaction_level'])
plt.xlabel('Number of Projects')
plt.ylabel('Number of Employees')
plt.show()
time_spent=data.groupby('time_spend_company').count()
```

```
plt.bar(time_spent.index.values, time_spent['satisfaction_level'])

plt.xlabel('Number of Years Spend in Company')

plt.ylabel('Number of Employees')

plt.show()

features=['number_project','time_spend_company','Work_accident','left',
        'promotion_last_5years','Departments ','salary']

fig=plt.subplots(figsize=(10,15))

for i, j in enumerate(features):

            plt.subplot(4, 2, i+1)

            plt.subplots_adjust(hspace = 1.0)

            sns.countplot(x=j,data = data)

            plt.xticks(rotation=90)

            plt.title("No. of employee")

fig=plt.subplots(figsize=(10,15))

for i, j in enumerate(features):

            plt.subplot(4, 2, i+1)

            plt.subplots_adjust(hspace = 1.0)

            sns.countplot(x=j,data = data, hue='left')

            plt.xticks(rotation=90)

            plt.title("No. of employee")

from sklearn.cluster import KMeans

left_emp = data[['satisfaction_level', 'last_evaluation']][data.left == 1]

kmeans = KMeans(n_clusters = 3, random_state = 0).fit(left_emp)

left_emp['label'] = kmeans.labels_

plt.scatter(left_emp['satisfaction_level'],left_emp['last_evaluation'],
        c=left_emp['label'],cmap='Accent')

plt.xlabel('Satisfaction Level')

plt.ylabel('Last Evaluation')
```

```
plt.title('3 Clusters of employees who left')

plt.show()

from sklearn import preprocessing

le = preprocessing.LabelEncoder()

data['salary']=le.fit_transform(data['salary'])

data['Departments ']=le.fit_transform(data['Departments '])

X=data[['satisfaction_level', 'last_evaluation', 'number_project', 'average_montly_hours',
        'time_spend_company', 'Work_accident', 'promotion_last_5years', 'Departments ',
        'salary']]

 y=data['left']

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20,random_state=42)

from sklearn.ensemble import GradientBoostingClassifier

gb = GradientBoostingClassifier()

gb.fit(X_train, y_train)

y_pred = gb.predict(X_test)

from sklearn import metrics

print("Accuracy:",metrics.accuracy_score(y_test,y_pred))

print("Precision:",metrics.precision_score(y_test,y_pred))

print("Recall:",metrics.recall_score(y_test, y_pred))
```